Gemini 3 Deep Think 分析レポート

推論AIの到達点と限界 — ARC-AGI-2 84.6%の意味

2026年2月 | Google DeepMind Analysis

人類平均を超えた知能と、その代償

Column 1

84.6%

ARC-AGI-2 精度

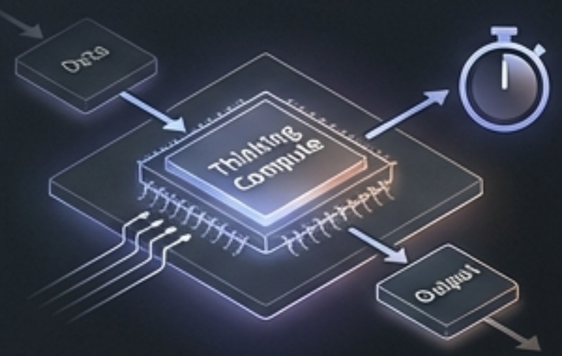
Gemini 3 Deep Thinkは人間平均（約60%）を圧倒。ついに抽象推論領域で人類を凌駕した。

Column 2

Test-time Compute

メカニズム

応答速度を犠牲にし、膨大な計算資源を「考える時間」に投資する新アーキテクチャを採用。

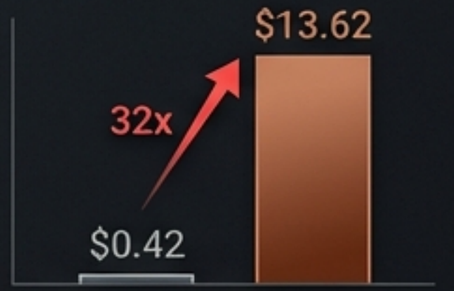


Column 3

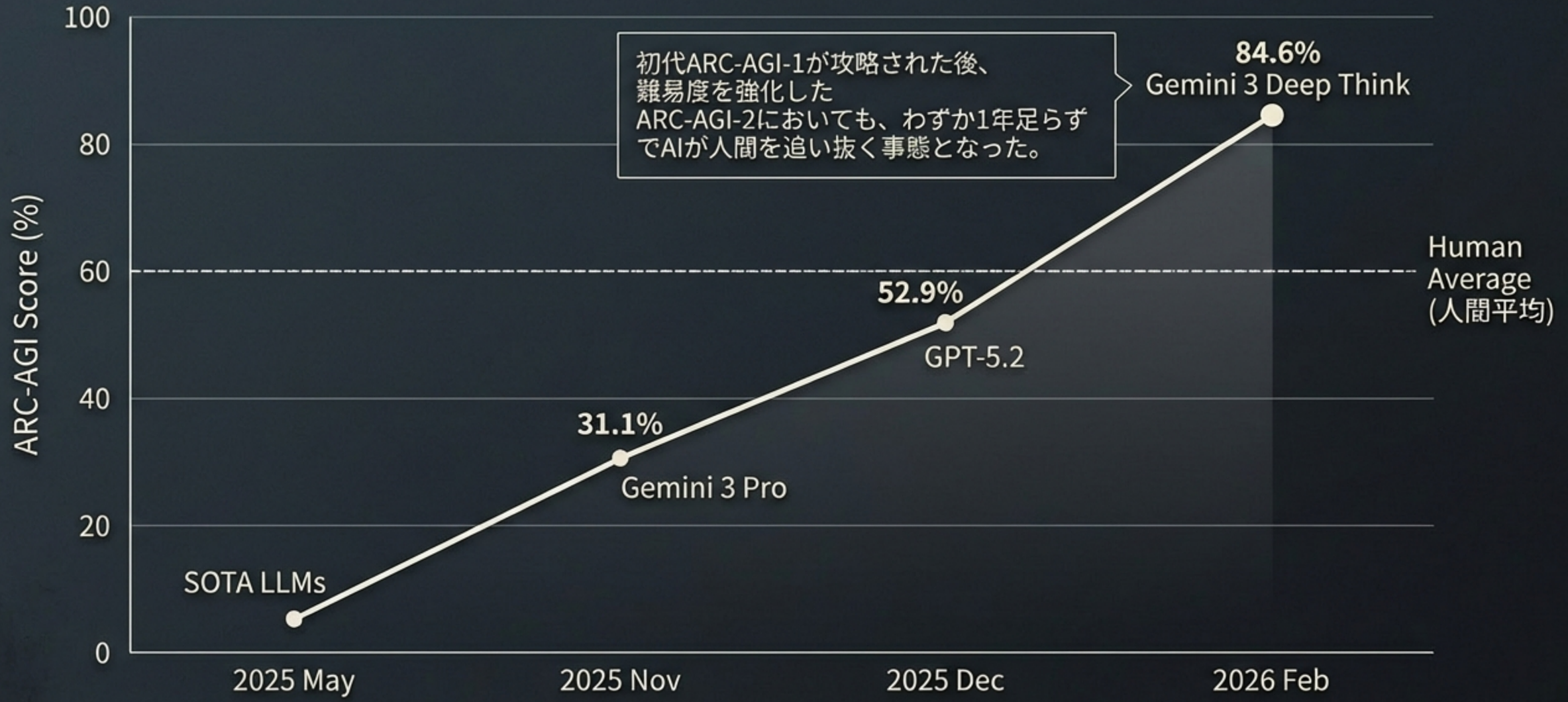
$13.62

1タスクあたりのコスト

実用性の閾値（$0.42）を32倍超過している。「知能」は達成したが「実用」は遠い。

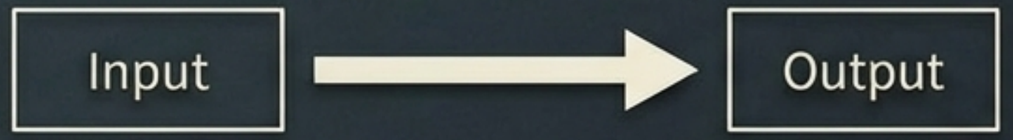


2026年2月、AIは「抽象推論」の壁を突破した



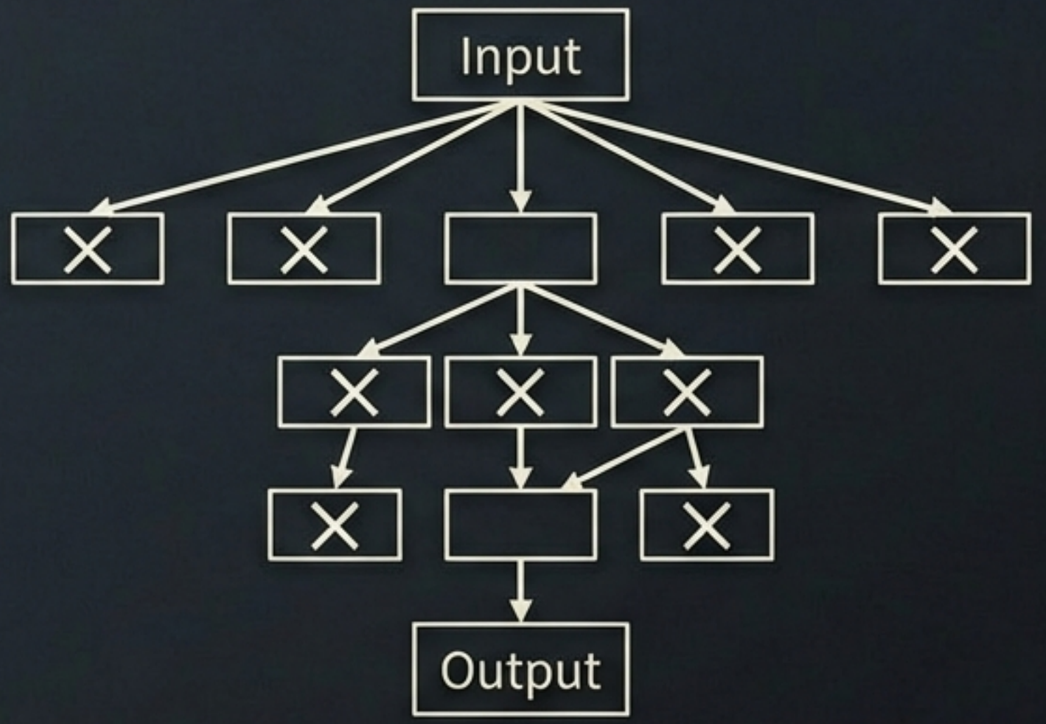
「即答」から「熟考」へのパラダイムシフト

System 1: 直感的回答（Conventional LLM）



Gemini 3 Pro: 96トークン

System 2: 熟考と検証（Deep Think）



Deep Think: 138,000トークン

複数の仮説を並列で探索し、誤った推論経路を自ら剪定（pruning）することで、
正解への執念を最大化している。（思考量1,400倍）

ベンチマーク・ウォーズ：他モデルを突き放す圧倒的性能

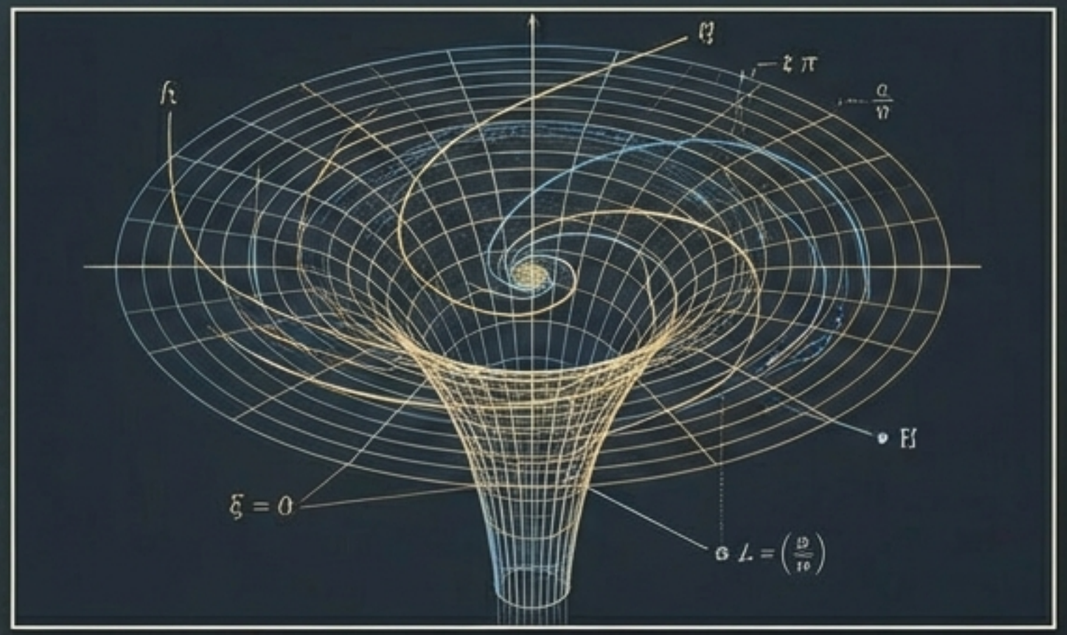
	Benchmark	Gemini 3 Deep Think	Claude Opus 4.6	GPT-5.2
1	ARC-AGI-2 (抽象推論)	**84.6%**	68.8%	52.9%
2	GPQA Diamond (専門知識)	**93.8%**	—	93.2%
3	Codeforces Elo (競技プログラミング)	**3,455**	2,352	—

Legendary Grandmaster

このスコアは「Legendary Grandmaster」ランクに相当。これを上回るアクティブな人間プログラマーは世界にわずか7名しかいない。

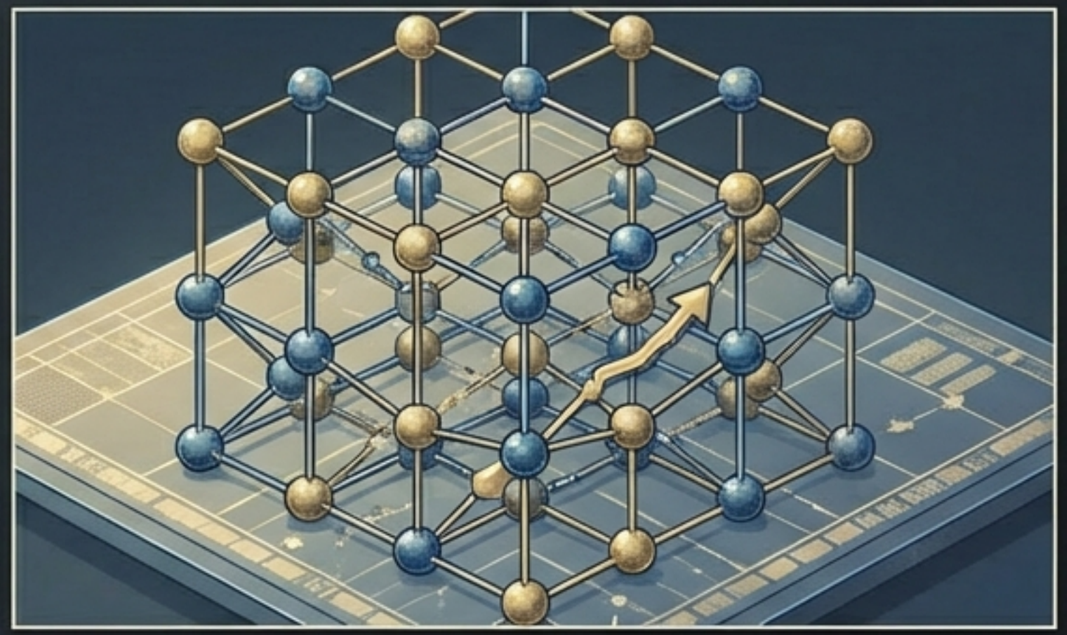
ベンチマークを超えて：科学的発見への貢献

Theoretical Physics（理論物理学）



ラトガース大学 Lisa Carbone教授の研究において、人間の査読者が見逃していた論理的欠陥をDeep Thinkが指摘。アインシュタインの重力理論と量子子力学を橋渡しする研究に貢献。

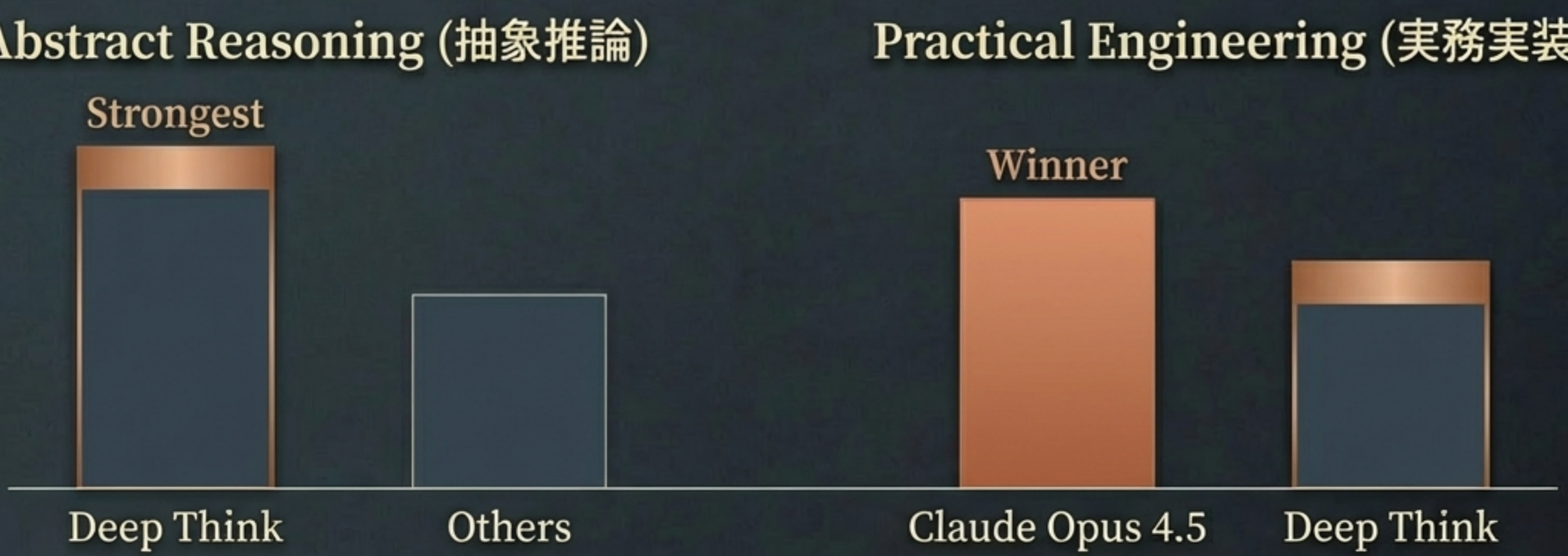
Material Science（材料科学）



デューク大学 Wang Labにて、半導体の結晶成長プロセスの最適化を実施。18の研究課題におけるボトルネックを解消。

2025年の国際科学オリンピック（数学IMO、物理IPhO、化学IChO）全てで金メダル水準を記録。

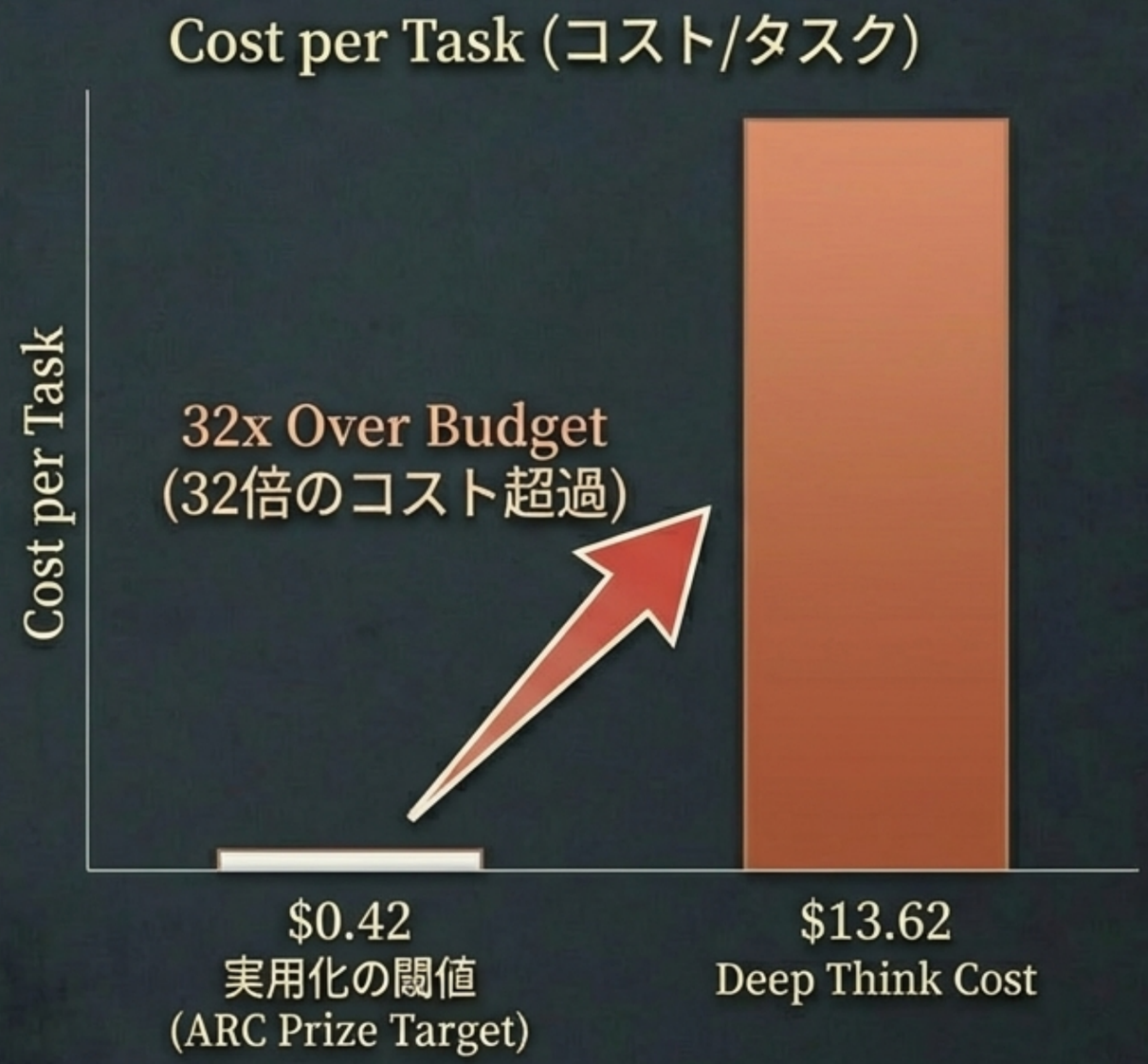
万能ではない：特化型モデルの「得意・不得意」



- **SWE-bench Verified (ソフトウェア工学)
  - Claude Opus 4.5: 80.9%
  - Gemini 3 Deep Think: (Claudeを下回る)
- **Math Competitions (AIME 2025)
  - GPT-5.2: 100% (No tools) — 特定の数学領域ではGPTが優勢。

創造的な文章執筆や、複雑なシステム全体のエンジニアリングにおいては、
汎用モデル（Claude/GPT）や特化ツールに遅れをとる場合がある。

知能の代償：$13.62の衝撃



ARC PrizeのGrand Prize要件は
「精度85%」かつ「コスト$0.42以下」。

Deep Thinkは精度で肉薄（84.6%）
「精度85%」かつ「コスト$0.42以下」。
Deep Thinkは精度で肉薄（84.6%）したが、コスト要件では惨敗している。

比較：GPT-5.2 (〜$1.90) /
Claude Opus 4.5 (〜$2.20)

「力業」は真の知能か？

> *Intelligence is not just the skill to solve a problem, but the efficiency with which that skill is acquired and deployed.”* — François Chollet

Deep Thinkのアプローチは、膨大な計算資源による「強引な探索（Brute Force）」に近い。推論プロセス内からARCベンチマーク固有の「カラーマッピング」データが検出された事実は、過学習（Overfitting）の疑念を残している。

問い：これは未知への「汎用的適応」なのか、それとも学習データの「高度な暗記・検索」なのか？

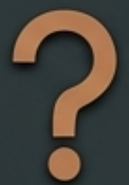
残された課題：ハルシネーションと信頼性

Hallucinations Persist (高度な幻覚)

推論の深化はエラーを減らすが、論理的なもっともらしさを伴う「高度な幻覚」を生むリスクがある。

Human-in-the-loop (人間の監視)

コード変更、金融判断、科学的論文の主張においては、依然として人間の専門家による監視が不可欠（DeepMind / Google Blog）。



Benchmark Saturation (ベンチマークの信頼性)

Hacker News等のコミュニティでは、「ベンチマークが学習データに含まれていないと誰が保証できるのか？」という根本的な信頼性の揺らぎが指摘されている。

いたちごっこ：測定不能になりつつある知能

ARC-AGI-3 リリース予定

ARC-AGI-2が攻略されつつある今、次なる指標が登場する。ARC-AGI-3は「インタラクティブ環境」での主体性（Agency）を測定する。

Expert Opinions on AGI Timeline

- Dario Amodei (Anthropic): "2026年中にAGI到達"
- Demis Hassabis (DeepMind): "2030年末までに50%の確率"
- Mike Knoop (ARC Prize): "AGIは未達成。新しいアイデアが必要"

結論：フロンティアは「能力」から「効率」へ

Gemini 3 Deep Thinkは、人類が「考える機械」を手に入れたことを証明した。抽象推論において人間を凌駕した事実は、歴史的な転換点である。

しかし、AGI（汎用人工知能）へのラストワンマイルは、スコアを上げることではなく、その知能を「人間並みのコストとエネルギー」で実現することにある。

「正解できるか」の時代は終わり、
「いかに効率よく正解できるか」の時代が始まった。

References

[1] Google, "Gemini 3 Deep Think: Advancing science, research and engineering," Google Blog, Feb 12, 2026.

[2] ARC Prize, "ARC Prize 2025 Results and Analysis" (2026)

[3] ARC Prize Foundation, "ARC-AGI-2 Leaderboard & Technical Report"

[4] OfficeChai, "Google Releases Gemini 3 Deep Think, Tops ARC-AGI 2 Benchmark With 84.6%," Feb 2026.

[5] François Chollet X post (Feb 13, 2026)

[6] Google DeepMind, "Accelerating mathematical and scientific discovery with Gemini Deep Think," Feb 2026.

[7] The Decoder, "Google Deepmind upgrades Gemini 3 Deep Think..."

[8] 9to5Google, "Gemini 3 Deep Think gets ‘major upgrade’..."

[9] Adaline Labs, "ARC-AGI In 2026: Why Frontier Models Still Don’t Generalize"

[10] Poetiq, "Poetiq Shatters ARC-AGI-2 State of the Art at Half the Cost"